Analysis of US Presidential Elections 1976-2020

MATH-014

By: Zoe Williams

Objective: The objective of this project is to examine the dataset "US elections." This dataset can be used to analyze the trends and patterns in presidential elections by exploring relationships between variables such as political parties, total votes, year, and location. Exploratory data analysis and data visualization will be used to interpret the data.

Introduction: This dataset contains 4287 columns and 15 rows. The variables collected were year, state, state po, state fips, state cen, state ic, office, candidate, party detailed, write in, candidate votes, total votes, version, notes, and party simplified. The NumPy library was used for numerical computation, the Pandas library was used for data manipulation and analysis, the matplotlib library was used for 2D plotting, and Seaborn was used for data analysis.

Method:

Data Cleaning: The quality of the dataset was satisfactory. Four rows contained missing values: candidate, party detailed, write in and notes. All of the rows with missing values were objects and therefore could not be filled in with the mean, median, mode, etc. The fillna function was used to fill the missing values with unknown. Due to the notes row missing all of its data the entire row was dropped. There were no duplicate columns.

Exploratory Data Analysis: Exploratory data analysis (EDA) was employed to analyze and investigate this dataset. Through manipulating this dataset, patterns and anomalies can be discovered. Additionally, the dataset's main characteristics can be summarized. The groupby and sum function was used to gather the total votes cast in each election year. This is the result:

	year
1976	605944064
1980	663902096
1984	609936856
1988	537099170
1992	770486377
1996	728343795
2000	783441739
2004	768259747
2008	992684830
2012	879479158
2016	941573717
2020	1865852281

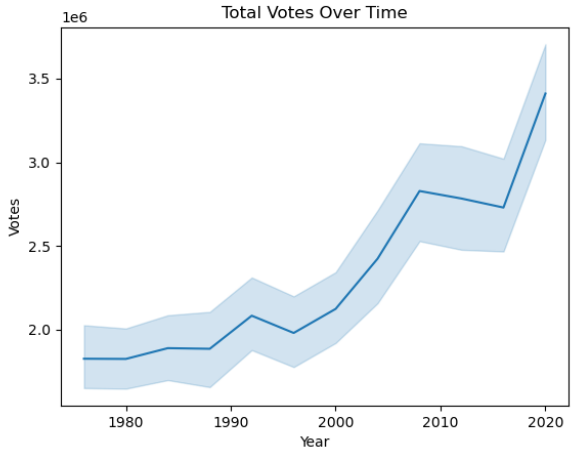
The same approach was used to collect state-wise total votes. The results showed that California, Florida, New York, Texas, and Illinois had the most votes among the states.

To determine the top candidates by total votes received, I selected the columns candidate and total votes. Then I used the sort.values function to rank them in ascending order. This is the result:

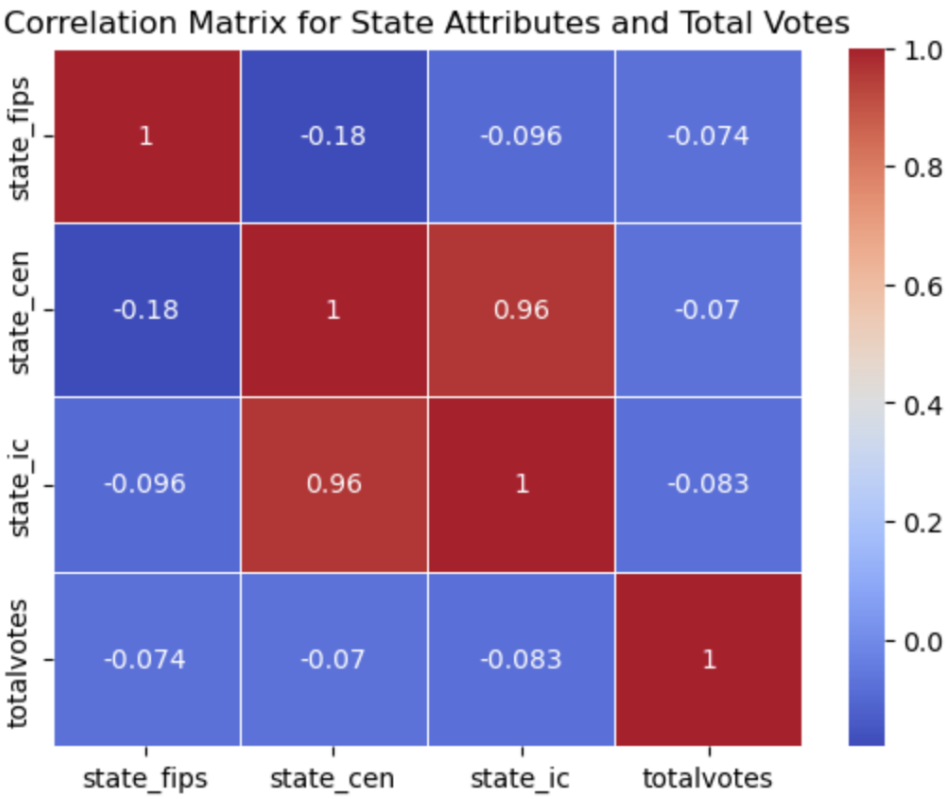
	candidate	totalvotes
3782	PIERCE, BROCK	17500881
3773	BIDEN, JOSEPH R. JR	17500881
3783	JOSEPH KISHORE	17500881
3781	MARK CHARLES	17500881
3780	JANOS, JAMES G. "JESSE VENTURA"	17500881
3779	CARROLL, BRIAN	17500881
3778	LA RIVA, GLORIA ESTELLA	17500881
3777	DE LA FUENTE, ROQUE ""ROCKY""	17500881
3776	HAWKINS, HOWIE	17500881
3775	JORGENSEN, JO	17500881
3774	TRUMP, DONALD J.	17500881
3430	LA RIVA, GLORIA ESTELLA	14181595
3427	JOHNSON, GARY	14181595
3426	TRUMP, DONALD J.	14181595
3425	CLINTON, HILLARY	14181595

Storytelling (Data Visualization and Interpretation):

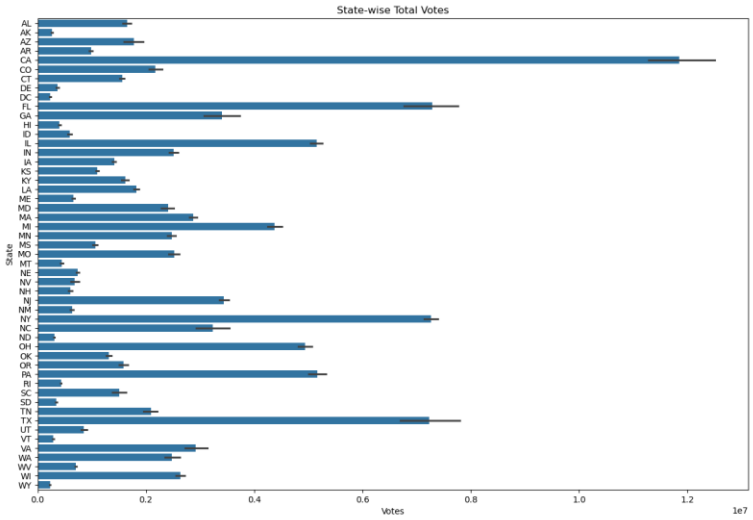
Total Votes Over Time: Seaborn was used to create a line graph of the total votes over time. Year was selected to be x-axis, and total votes were selected to be y-axis. Although the graph is not a perfect linear line, year and total votes exhibit a positive relationship. There are multiple dips over time, such as 1993-1996 and 2008 to 2012, but there are strong increases from 1995 to 2008 and 2015 to 2020. These increases can be attributed to efforts to fight voter suppression and grassroots organizations that aim to increase voter turnout, among other variables.



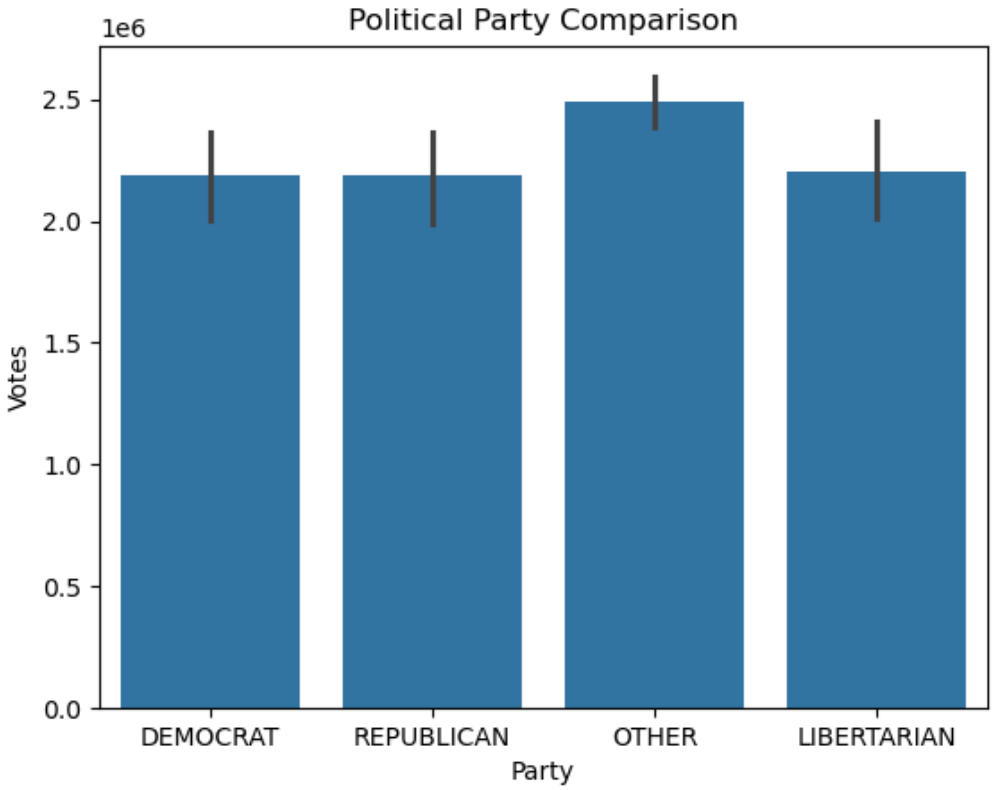
Correlation Between State Attributes and Total Votes: To determine the correlation between state attributes and total votes, seaborn was used to create a heatmap. total votes, and all variables used for location (state po, state fips, state cen, state ic) were selected. The heatmap demonstrates that there is little to no correlation between state attributes and total votes. On the plot, total votes have negative values and blue boxes for all of that state attribute variables. This could allude to location having no impact on total votes.



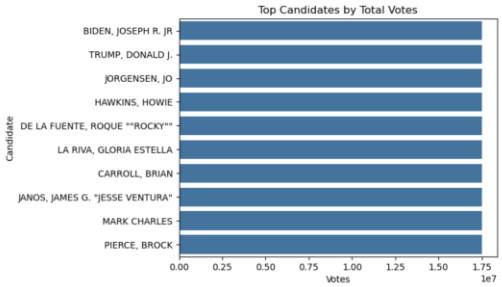
State-wise Total Votes: To display states wise total votes, a horizontal bar graph was created using seaborn. Total votes were selected to be the x-axis and states was the y-axis. This graph shows that the states with the highest votes are California, Texas, and Florida. The states with the lowest votes are Wyoming, Alaska, and Vermont. These results show that population is strongly correlated with total votes.



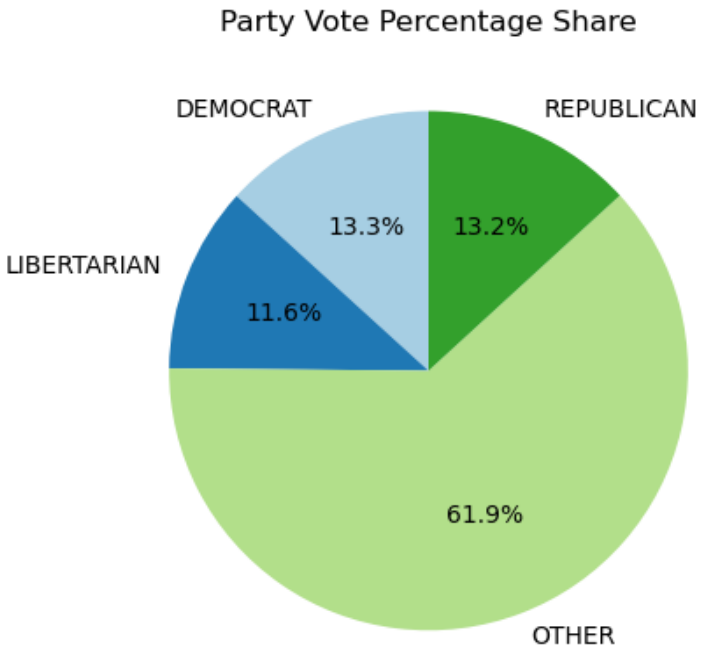
Party Performance Comparison: Seaborn was used to create a bar graph comparing political parties. I selected party simplified for the x-axis and total votes for the y-axis. This graph demonstrates that democrat, republican, and libertarian all roughly received the same number of votes. Other, is noticeably higher than the other parties, however this variable contains a large variety of parties. This graph alludes to there being no dominant or more successful party in the two-party system.



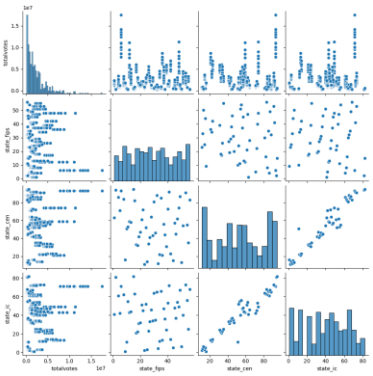
Top Candidates by Total Votes: Seaborn was used to create a bar graph of top candidates. Total votes were used for the x-axis and the top ten candidates were used for the y-axis. The graph shows that all of the top ten candidates roughly received the same number of votes. This could possibly be the number when 270 electoral votes were met or a mistake in the dataset. This bar graph shows an anomaly in the dataset.



Party Vote Percentage Share: Matplotlib.pyplot was used to create a pie chart of the party vote percentage share. The pie chart shows that democrats received 13.3%, republicans received 13.2%, libertarians received 11.6%, and other political parties hold 61.9% of the vote. These numbers show that the party's within the two-party system hold nearly the same amount of votes, while smaller political parties dominate a large share of total votes.

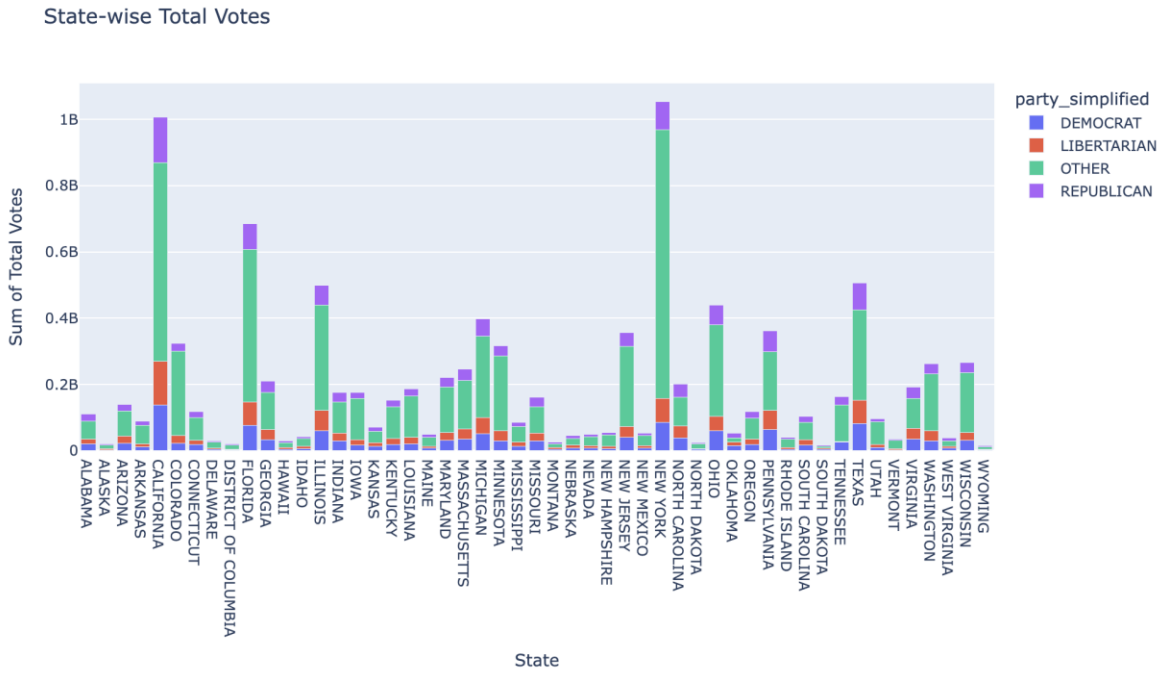


Correlation Between State Attributes and Total Votes: Seaborn was used to create a pairplot of state attributes and total votes. The state attributes and the total votes scatter plots appear clustered near the y-axis or x-axis. This alludes to a weak correlation between the variables.

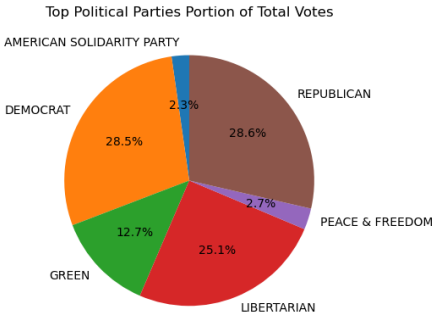


State-wise Total Votes and Political Party: This graph displays the party breakdown for each state. The bar graph clearly shows which states have high voter turnout; however, it is still hard

to determine red and blue states due to the impact of the other party category. The green color is the other party category, and it makes up the majority of many of the state's bars, for most states it appears that republican, democrats and libertarian are around the same size.



Party Vote Percentage Share: Due to the overwhelming amount of total vote percentage the other political parties received; my additional graph is a pie chart breaking down some of the parties. The top three identified parties were selected, along with the top three parties from the other category. This pie chart helps display the true portion of the votes received by non-mainstream parties. This graph is helpful because the previous pie chart exaggerates the influence the other parties had on elections, while this pie chart demonstrates that impact is not as big as it is perceived.



Conclusion: Through data visualization and exploratory data analysis (EDA), this dataset has provided crucial information on U.S. Presidential elections from 1976 to 2020. By examining trends over time, correlations between state attributes and total votes, and the influence of

political parties and candidates, we were able to gain a deeper understanding of the electoral process. One of the most prominent findings was the overall growth in total votes, particularly in the years following 1992, with a sharp increase in voter turnout seen in the 2020 election. The analysis of state-wise total votes revealed that states like California, Texas, and Florida contributed the most to the total vote count, emphasizing the role of population size in shaping election outcomes. Regarding political parties, the visualizations suggested that while the major political parties (Democrat, Republican, and Libertarian) received similar levels of support, the "Other" category (comprising smaller parties) held a significant share of the total vote. Furthermore, the analysis of top candidates by vote share illustrated an anomaly, where many candidates appeared to have received identical vote totals. In conclusion, this dataset reveals important trends, such as the growth of total votes and the dominance of large states, it also raises questions about the role of minor political parties and anomalies in candidate vote counts.

References: Dataset is from kaggle: https://www.kaggle.com/datasets/tunguz/us-elections-dataset

Acknowledgements: Dr. Nerolu for introducing me to the world of data science!